5BZCAT BZUs的分类*

朱惊天，樊军辉，蔡金庭

(广州大学天体物理中心，广东 广州，510006)

(广东省高校天文观测与技术重点实验室，广东 广州，510006)

(广州天文观测与技术重点实验室，广东 广州，510006)}

摘 要：本文用4种机器学习方法：支持向量机(SVM)、随机森林(RF)、集成学习(EM)和多层感知机(MLP)，将5BZCAT中227个BZUs分为BL Lacs候选体和FSRQs候选体，并通过特征工程和网格搜索方法提高分类准确率。综合4种分类器的分类结果，本文中，将判别概率阈值设为0.8，得到33个BL Lacs候选体和119个FSRQs候选体。

关键词：耀变体；蝎虎天体；平谱射电类星体；机器学习；分类

中图分类号：TN216　　**文献标识码**：A　　**文章编号**：1007-2276-(2004)4-0338-05

耀变体(blazar)是活动星系核(AGNs)的一个特殊子类。由于其相对论喷流与观测者视线夹角较小，它们表现出极端的观测性质：快速且大幅度的光变，高光度，高偏振，视超光速运动，强烈的高能γ-ray辐射等。根据光学发射线强弱，耀变体通常分为蝎虎天体(BL Lacs)和平谱射电类星体(FSRQs)，BL Lacs只有很弱的发射线，甚至没有发射线，而FSRQs有强发射线[1-4]。耀变体的能谱分布(Spectral Energy Distribution，SED)呈现双峰结构，根据同步峰频($\log\nu_p^s$)大小，又可以将耀变体分为低峰频(LSP)、中峰频(ISP)、高峰频(HSP)耀变体[5-6]。由于很多耀变体缺乏光学光谱信息，不能有效确定它们是BL Lacs还是FSRQs。这种类别不确定的耀变体在不同的文献中有不同的名称，例如，在Roma-BZCAT 5期源表[7]中被称为BZUs (Blazars of Uncertain type)；而在Fermi/LAT 源表(3FGL，4FGL)[8][9]中被称为BCUs(Blazar Candidate of Uncertain types)。

5BZCAT[7]、3FGL[8]和4FGL[9]源表均包含了多于一千个的耀变体，以及它们的红移、同步峰频、多波段流量/流量密度、多波段有效谱指数等观测数据。这些源表为研究耀变体的性质提供了大样本。同时，这些源表中也包含了至少数百个BCUs/BZUs。对BCUs的分类已经引起不少作者的兴趣，随着机器学习(Machine Learning，ML)方法在天文领域的广泛应用[10-12]，很多BCUs的分类工作也使用了这些方法[13-19]。例如，Fermi/LAT 3期AGN源表(3LAC)[20]中的高置信度样本 (3LAC Clean Sample)共有402个BCUs，Kang等[15]对其中的无缺失数据的400个BCUs进行了分类。他们用了4种ML分类方法，综合这些分类器的结果，获得了246

*基金项目：国家自然科学基金(11733001、U1531245)；广东省自然科学基金(2017A030313011；2019B030302001)；广东省和广州市重点学科资助.

作者简介：朱惊天，男，硕士，研究方向：活动星系核的数据处理。

个BL Lacs和候选体74个FSRQs候选体；Fermi /LAT 4期源表(4FGL)中共有1312个BCUs，Kang等[16]用3种ML分类方法对它们进行分类，同时考虑3种分类方法的结果，得到724个BL Lacs候选体和332个FSRQs候选体，仍有256个BCUs没有给出明确的光学分类。为了对5BZCAT中的BZUs的光学分类进行评估，本文使用了支持向量机(SVM)、随机森林(RF)、集成学习(EM)和多层感知机(MLP)这4种ML分类方法，将BZUs分类为BL Lacs候选体和FSRQs候选体。本文结构如下：第1节介绍样本；第2节介绍分类方法；第3节给出分类结果与讨论；第4节是总结。

1 样本

5BZCAT中共有3561个耀变体，其中有1425个BL Lacs、1909个FSRQs和227个BZUs。从5BZCAT给出的参数中，除去坐标、源名等无效参数，选取8个可用参数，分别是：红移(z)，1.4GHz处的射电流量密度(f_{R})，射电–光学有效谱指数(α_{RO})，X-ray–光学有效谱指数(α_{OX})，射电–X-ray有效谱指数(α_{RX})，光学R波段视星等(m_{R}')，0.1-2.4keV的X-ray 积分流量(F_{X})，1-100GeV的γ-ray光子积分流量($F_{\gamma}^{\mathrm{pho}}$)。还从NED(NASA/IPAC Extragalactic Database[1]) 中获得R波段的消光系数(A_{λ})，对 m_{R}' 做了消光改正，得到R波段真实视星等(m_{R})。最终，本文中，用 z， α_{RO}， α_{OX}， α_{RX}， f_{R}， m_{R}， F_{X}， $F_{\gamma}^{\mathrm{pho}}$ 共8个参数作为分类所需的数据集特征。

2 分类方法

ML是人工智能领域中一种新兴的方法，其包含多种分类模型(分类器)和回归模型，这些模型能从已知数据中学到某种规律，并应用到新数据中。ML方法在天文领域的分类和回归研究中有着良好表现[10-12]。Scikit−learn(sklearn)[21]是python提供的ML模块，其中包含了许多ML算法，例如数据预处理方法和多种ML分类器。分类器是通过学习已知类别的数据，获得分类标准，然后用于未知类别的数据。通常已知类别的数据将按一定比例随机划分为训练集和测试集，未知类别的数据则作为预测集。训练集用来训练分类器，在分类过程中学习训练集的参数蕴含的信息，确定不同类别的区分标准；测试集则用来测试分类器的性能；利用优化分类模型（标准），评估预测集的分类。

样本中，未知类别的227个BZUs作为预测集。已知类别的数据3334个耀变体（1425个BL Lacs和1909个FSRQs）利用klearn.train_test_split函数将其按7：3的比例随机划分为训练集

[1] http;//ned.ipac.caltech.edu/

和测试集。每次划分训练集和测试集时，为确保训练集和测试集中的BL Lacs 和FSRQs 的数量比例与样本相同，设置随机种子为固定值（如，random_state=1）。文中，训练集有2333个耀变体（997个BL Lacs和1336个FSRQs）；测试集有1001个耀变体（428个BL Lacs和573个FSRQs）；为了确保结果的稳定性，对sklearn.train_test_split函数中random_state(随机数种子)取5个不同值：0、1、2、3、4，用这5个随机数随机划分训练集和测试集，得到5个不同的训练集和对应的测试集：训练集1、测试集1，... 训练集5、测试集5。本文中，在5个训练集上分别训练分类器，得到5个不同的候选分类器，利用5个测试集上测试5个候选分类器的性能，然后选择性能最优的1个，用于预测227个BZUs(预测集)的分类。

2.1 分类器

4种分类器：支持向量机(Support Vector Machine，SVM)、随机森林(Random Forest，RF)、多层感知机(Multi-Layer Perceptron，MLP)、和集成学习(Ensemble Learning，EM) 的介绍如下：

(1)对于线性可分的两类样本，可以在参数空间中找到无穷多个超平面，将两类样本分隔在超平面两侧，其中距离超平面最近的样本点被称为支持向量(Support Vector，SV)。SVM的原理是寻找唯一的最优超平面，使得SV到该最优超平面的距离最大化。如果两类样本是非线性可分的，SVM可以将样本映射到高维(甚至无穷维)空间中，然后寻找高维空间中的最优超平面。

(2)决策树(Decision tree，DT)的结构是二叉树，分类时，信息进入节点时进行二元判断，当一个节点无法判断出类别，则分裂为二个，直至判断出类别为止。由于DT容易陷入节点过度分裂的情况，导致分类器泛化性差。而随机森林(RF)由大量DT构成，其中DT之间相互独立，RF随机划分训练集和参数给每个DT，分类结果由所有DT投票决定，RF的*泛化性能*往往优于单个DT；

(3)MLP 是人工神经网络（Artificial Neural Network，ANN）的一种。ANN 是一系列模仿生物神经网络(如人类大脑)结构的算法。这些结构由多个人工神经层组成，包括一个输入层、一个或多个隐藏层和一个输出层。每一个人工神经层可以识别数据中的特定元素，然后将结果传播到下一人工神经层。通过综合每一个神经层的结果，ANN 可以学习识别数据中的复杂特性。

(4)EM通过某种集成规则，将一组基评估器的结果集成，其性能往往优于单个基评估器。本文将SVM，RF，MLP这3种分类器做为EM的基评估器，集成规则为软投票，即给每个基评估器：SVM，RF，MLP输出的类别概率一个权重，权重在[0，1]区间。然后对基评估器的类别概率求加权和，作为EM输出的类别概率。本文尝试了多种权重组合，并选取其中最优的一个。

2.2 性能指标

ML常用的性能指标有准确率(accruacy)，精准率(precision)，召回率(recall)等。本文中，只考虑准确率：

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{FP} + \text{FN} + \text{TP} + \text{TN}}$$

其中，TP (True Positive)是被正确分类的正类别样本点数；TN (True Negative)是被正确分类的负类别样本点数；FP (False Positive)是被误分类的正类别样本点数；FN(False Negative)是被误分类的负类别样本点数。准确率(accuracy)代表被正确分类的样本点数占样本容量的比重。

2.3 数据集转换和特征选择

ML中，描述样本点属性的参数被称为特征，本文中，分类所用的特征即是5BZCAT中的8个参数(见第1节)。真实样本的特征往往还包含缺失值、噪声、无关信息、冗余信息等，它们会影响ML分类器的性能；因此，本文中，在使用ML分类器之前，需要先对原始数据做数据集转换和特征选择的处理，这样做的目的是保证最大限度从原始数据中提取有效特征供ML分类器学习。数据集转换通常包括预处理数据和无监督降维，特征选择和无监督降维都是降维方法，可以减少特征。分别对数据集转换和无监督降维做如下概述：

(1)预处理数据一般包括缺失值补全和标准化。本文中，若某个特征有缺失值本文中，用同类特征的平均值填充；而标准化是将所有特征映射到到相同区间，以免某些特征的量级比其他特征小，导致分类器只学习量级大的特征。本文中，用sklearn中的preprocessing.StandardScaler将所有特征化为标准正太分布；

(2)降维一般包括特征选择(Feature Selection，FS)和无监督降维(unsupervised dimensionality reduction，UDR)，其可以降低特征维度，减少计算成本，并能提升分类器性能。本文中，采用的FS和UDR分别为序列向后选择(Sequential Backward Selection，SBS)和主成分分析(Principal Component Analysis，PCA)。其中，SBS选择原特征集的子集，而PCA将原特征映射到新空间中，再选取新特征集的子集。对SBS和PCA简单介绍如下：SBS不断从当前全部特征中舍去一个特征，直到所剩特征数量满足要求，每次被舍弃的特征满足：与舍弃其他特征相比，舍弃该特征后分类器的性能损失达到最小；而PCA不依赖分类器，它将样本点从原n维特征空间映射到新的n维正交空间，得到n个两两线性无关的新特征，新空间中，每个坐标轴被称为主成分(Principal Component，PC)，在每个PC方向上，样本点的分离都达到最大。第1个PC代表样本方差最大的方向，称为第1主成分，其余坐标轴称为第2，第3，...，第n主成分，每个主成分均为n个原特征的线性组合。它们对样本方差的贡献率依次递减，可根据需要取前k个主成分，$k \le n$。

本文中，对SVM和RF都做了SBS和PCA，并将分类准确率与不做降维的分类准确率进行比较。对于SBS，本文中，在8维原特征空间的训练集1上用SBS筛选特征。SBS每次减少1

个特征直至只剩1个特征，在此过程中观察分类器在不同维度的特征空间中的性能，选出最优的特征空间；而主成分共有8个，本文中，舍弃第8主成分，其方差贡献率只有0.0305%，其余主成分的方差贡献率均大于5%。SBS和PCA的结果分别如图1，图2所示。图1为SVM和RF的SBS结果，横坐标为特征数量，纵坐标为对应的分类器准确率。图2为PCA结果，横坐标为各主成分，纵坐标为对应的方差贡献率，为了便于观察，图中第8主成分的方差贡献率被放大了50倍。而MLP的一个优势是无需做太多特征工程，因为ANN的隐藏层能自动提取有效特征，并能自适应特征间的非线性关系，因此本文中，没有对MLP做数据降维。对于EM，本文中，在每个训练集上将性能最优的SVM，RF，MLP分类器以最优的权重集成。

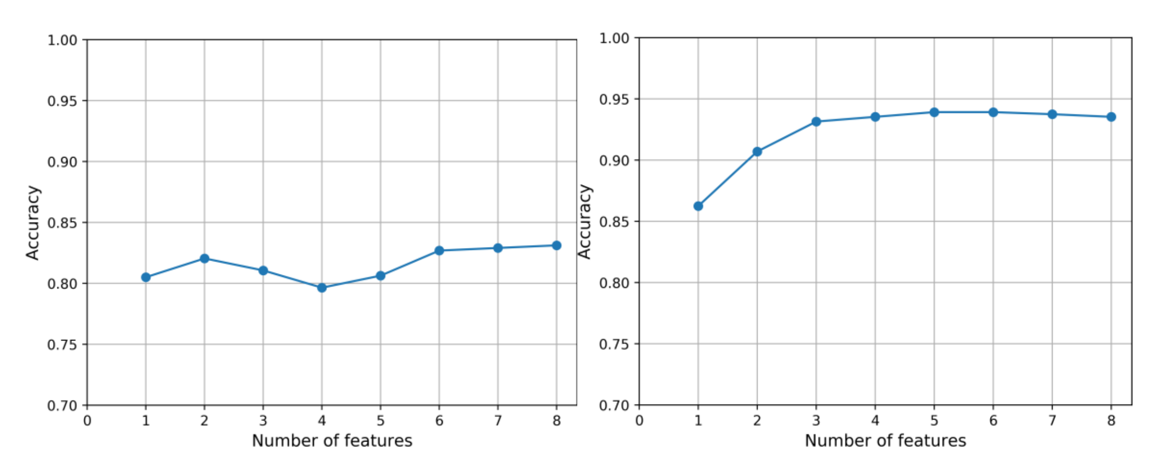


图 1　SBS结果图。左图为SVM的SBS结果，右图为RF的SBS结果

Fig.1　The result graph of SBS. Left:SBS for SVM;　Right: SBS for RF

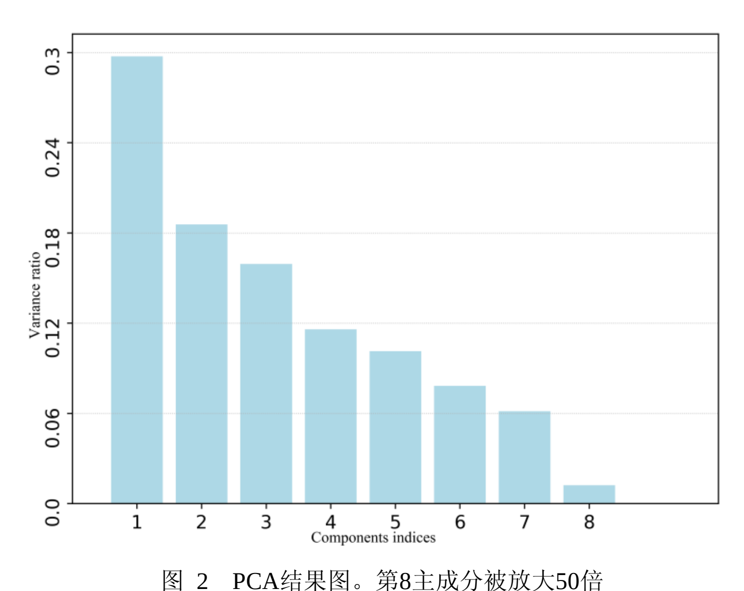


图 2　PCA结果图。第8主成分被放大50倍

Fig.2　The result graph of PCA. The 8th principal component is magnified by 50 times

2.4 **超参数**

在sklearn提供的ML和分类器中，有部分函数参数属于自由变量，其被称为超参数(Hyper Parameters，HPs)。HPs不能通过训练分类器得到，而要在训练分类器前人为赋值。对于HPs

的取值，本文中，使用网格搜索方法(Grid Search，GS)，找出最优值。具体来说，本文中，指定一组候选值，GS以暴力穷举的方式，选出能最大化分类器准确率的值。本文中，对四种分类器中较重要的HPs使用了GS，例如SVM中的C和MLP中的alpha，这两个HPs可以提高模型的泛化性能。

3 结果与讨论

4种分类器在测试集上的准确率和BZUs的分类结果概述如下：

(1)对于SVM，分类结果显示，训练集3上，8维原特征空间中的分类器准确率最高，为84.62%（见表1，图1）。在所有训练集上，PCA选7维主成分空间中的分类器准确率均不如其他特征空间；

(2)对于RF，SBS在训练集1上选出了5个最优特征：z，$\alpha_{\rm RO}$，$\alpha_{\rm OX}$，$m_{\rm R}$，$F_{\gamma}^{\rm pho}$。分类结果显示，训练集5上，SBS选5维子特征空间中的分类器准确率最高，为94.41%。在所有训练集上，SBS选5维子特征空间中的分类器准确率均优于在8维原特征空间中的，而PCA选7维主成分空间中的分类器准确率均不如其他特征空间；

(3)对于MLP，分类结果显示，在训练集5上，8维原特征空间中的分类器准确率最高，为94.21%；

(4)对于EM，在每个训练集上，本文中，选取准确率最高的SVM、RF、MLP分类器集成。分类结果显示，在训练集5上，分类器准确率最高，为94.81%（见表1，图1）。此时成员分类器SVM，RF，MLP的权重分别为：0，0.62，0.38。

分类器在5个测试集上的准确率表明，对于同一种分类器和同一个特征空间，5个测试集上分类器的准确率都相近，**这**说明本文中，的分类结果是稳定的；并且对于RF，**当**SBS将8个特征减少到5个时，反而最大程度提高了分类器准确率。相反地，SVM和RF使用了PCA后性能均明显下降，这可能是由于原特征间有非线性关系，不能很好地被分离成两两线线性无关的新特征。分类结果显示，4种分类器中，SVM的准确率明显低于其余3种。原因可能是在8参数空间中，BL Lacs和FSRQs不能很好地被线性边界分开；而RF、MLP和EM分类器能很好地捕捉到非线性分类边界，因此其性能都较良好且优于SVM。4种分类器在5个测试集上的准确率和GS确定的最优HPs分别展示在表1和表2中，表1中各列说明如下，第1列为分类器名称；第2列为测试集名称； 第3-5列分别为8维原特征空间、SBS选特征空间、7维主成分空间中的分类器准确率；第6列为EM分类器准确率。表2中各列说明如下，第1列为分类器名称； 第2列为测试集名称；第3-7列分别为8维原特征空间、SBS选特征空间、7维主成分空间中分类器的最优HPs。本文中，同样在图1中展示了4种分类器的准确率，可以更直观地看到每种分类器在每个训练集上的准确率。图1中4张子图的横轴均为测试集名称，纵轴为分类器的准确率； 图1上半部分两张子图，从左到右分别为SVM和RF分类器，其中蓝色、橘

色、绿色柱状图分别代表8维原特征空间、SBS选特征空间、7维主成分空间中分类器的准确率； 图1下半部分两张子图，从左到右分别为MLP分类器在8维原特征空间和EM分类器的准确率。

本文中，选择在测试集上准确率最高的4个分类器，用它们对227个BZUs进行分类，得到每个BZU的 $p_{\rm BL\ Lacs}$。若将判别概率的阈值设为； $p_0 = 0.5$，即某个源的 $p_{\rm BL\ Lacs} > 0.5$ 则判为BL Lacs，否则判为FSRQs。则SVM、RF、MLP、EM分别给出116、106、112、112个BL Lacs候选体和111、121、115、115个FSRQs候选体。本文中，将4种分类器的分类结果与3FGL、4FGL和Kang等[15-16]中的BL Lacs和FSRQs进行了比较，发现本文中，的分类结果与其他文献并不完全一致，例如，对于EM的分类结果，分别有8、10、9、14个BZUs的分类与3FGL、4FGL和Kang等[15-16]的分类不同。本文中，尝试进一步改进分类方法，以求减少与其他文献分类不一致的BZUs(不匹配源或mismatched BZUs)数量。本文尝试了两种改进方法：(1)对 p_0 分别取0.5、0.6、0.7、0.8、0.9、0.95这6个不同值，并比较分别取6个值时4个分类器的mismatched BZUs数量，即对 p_0 做GS。比较结果显示，当 $p_0 = 0.7$ 和 $p_0 = 0.8$，与3FGL对比分类结果时，SVM和RF的mismatched BZUs数量明显下降。其余情况下，mismatched BZUs数量随 p_0 取值不同没有显著变化；(2)对于某个BZU的预测类别，本文中，同时考虑4个分类器的分类结果，即只有当4个分类器的预测类别都一致时，才认为该BZU属于该预测类别，否则认为该BZU的类别是不确定的(unknowns， unks)。即对于某个源，只有当4个分类器同时预测其类别为

表 1 ML分类器性能

Table 1 Accuracy for ML classifiers

Model (1)	Test data (2)	Accuracy(8 features) (3)	Accuracy(SBS) (5)	Accuracy(PCA) (7)	Accuracy(EM) (9)
SVM	Test data1	0.8192	0.8192	0.7922	-
	Test data2	0.8272	0.8272	0.8002	-
	Test data3	0.8462	0.8462	0.8052	-
	Test data4	0.8192	0.8192	0.7902	-
	Test data5	0.8292	0.8292	0.7872	-
RF	Test data1	0.9211	0.9241	0.8761	-
	Test data2	0.9341	0.9421	0.8891	-
	Test data3	0.9381	0.9421	0.8971	-
	Test data4	0.9341	0.9341	0.8781	-
	Test data5	0.9331	0.9441	0.8821	-
MLP	Test data1	0.9271	-	-	-
	Test data2	0.9291	-	-	-
	Test data3	0.9191	-	-	-

	Test data4	0.9271	-	-	-
	Test data5	0.9421	-	-	-
EM	Test data1	-	-	-	0.9331
	Test data2	-	-	-	0.9461
	Test data3	-	-	-	0.9431
	Test data4	-	-	-	0.9391
	Test data5	-	-	-	0.9481

BL Lac或FSRQ时，本文才认为该源是BL Lac候选体或FSRQ候选体，否则认为该源的类别是unks。依此标准，再次比较当 p_0 取方法(1)中的6个不同值时，mismatched BZUs数量。此时的比较结果表明，6个不同 p_0 的mismatched BZUs数量相当，且均显著小于方法(1)中的mismatched BZUs数量。当 $p_0=0.8$ 和 $p_0=0.9$ 时，unks的数量最少。本文中，希望有尽可能多的BZUs被分类，综合以上2种分类改进方法，本文使用的分类改进方法是：p_0 取0.8时，由4个分类器共同决定每个BZUs的类别。

4个分类器的mismatched BZUs数量如表3和表4所示，p_0 取0.8时，227个BZUs由4个分类器共同决定的类别，以及和其他文献共同源的类别均展示在表5（附件）中。第1列代表进行分类比较的文献，第2列为不同文献中BL Lacs和FSRQs的总数，第3列为分类器名称，第4-9列为 p_0 取不同值时，4种分类器的mismatched BZUs数量；表4中，第1列代表进行分类比较

表 2　各分类器的最优超参数

Table 2　Optimal hyper parameters for ML classifiers

Model (1)	Test data (2)	HP (3)	GS(8 features) (4)	GS(SBS) (5)	GS(PCA) (6)	GS(EM) (7)
SVM	Test data1	C	100	100	100	-
	Test data2		1000	1000	1000	-
	Test data3		1000	1000	10	-
	Test data4		1000	1000	100	-
	Test data5		1000	1000	10	-
RF	Test data1	criterion, max_features, n_estimators, oob_score	entropy, None, 400, False	entropy, log2, 50, False	gini,log2, 400, False	-
	Test data2		entropy, None, 50, False	entropy, log2, 100, False	entropy, log2, 100, False	-
	Test data3		entropy, log2, 1000, False	entropy, log2, 400, False	gini, log2, 1000, False	-
	Test data4		entropy, None, 400, False	entropy, log2, 400, False	entropy, log2, 1000, False	-

	Test data5		entropy, None, 1000, False	gini, log2, 100, False	gini, None, 1000, False	-
MLP	Test data1	activation, alpha, hidden_layer_sizes, solver	tanh, 1, (30, 20, 10), lbfgs	-	-	-
	Test data2		tanh, 0.001, (20, 10, 5), lbfgs	-	-	-
	Test data3		tanh, 1, (50, 30, 10), lbfgs	-	-	-
	Test data4		tanh, 1e-05, (70, 50, 30),adam	-	-	-
	Test data5		tanh, 1, (50, 30, 10), lbfgs	-	-	-
EM	Test data1	weights	-	-	-	[0.13 0.44 0.43]
	Test data2		-	-	-	[0.08 0.87 0.05]
	Test data3		-	-	-	[0. 0.92 0.08]
	Test data4		-	-	-	[0.01 0.88 0.11]
	Test data5		-	-	-	[0. 0.62 0.38]

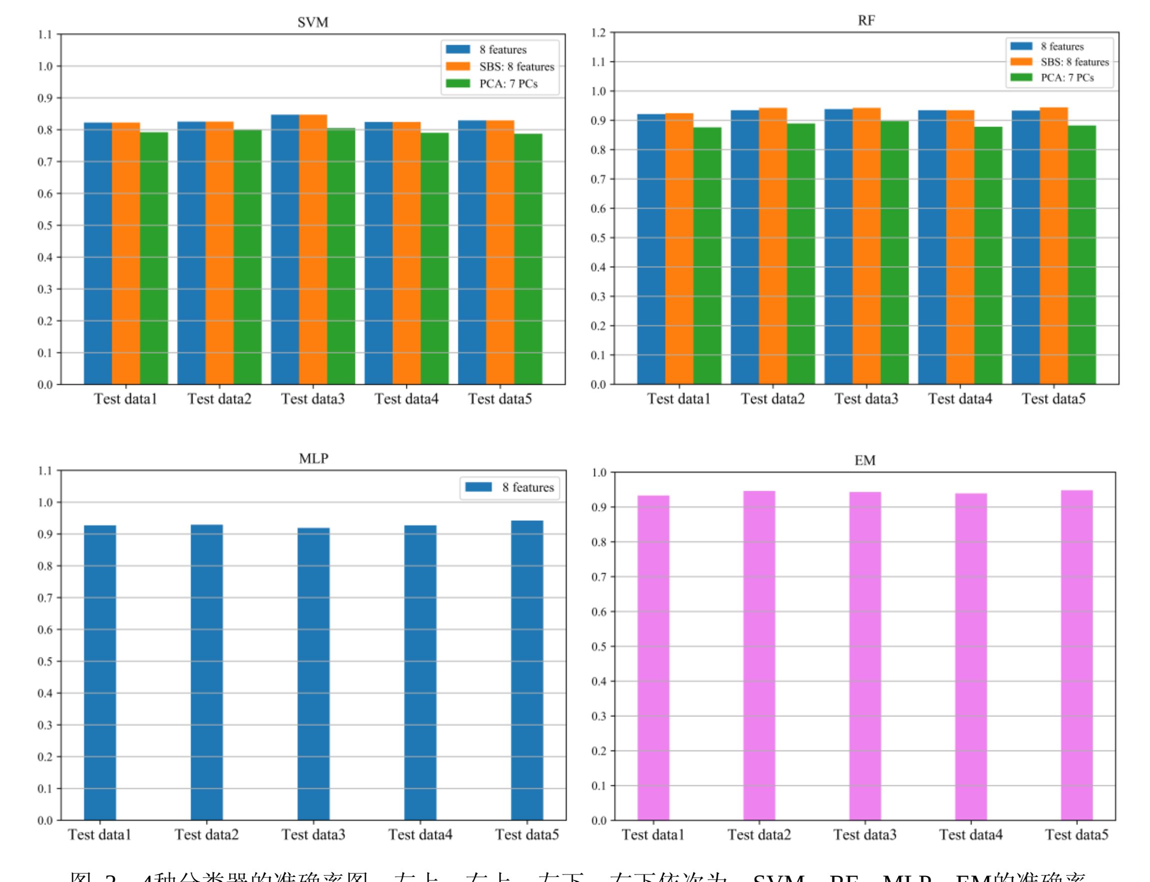


图 3　4种分类器的准确率图，左上、右上、左下、右下依次为：SVM、RF、MLP、EM的准确率

Fig.3　The accuracy graph of 4 classifiers，the upper left、upper right、lower left、lower right: accuracy for SVM，RF，MLP，EM

的文献和unks，第2列为不同文献中BL Lacs和FSRQs的总数，第3-8列为4个分类器共同决定BZUs 类别时，不同的 p_0 取值对应的 mismatched BZUs 和 unks 数量；表 5 中，第 1-3 列为 BZUS 在 5BZCAT、3FGL 和 4FGL 中的名称；第 4-7 列为 4 种分类器的预测的 $p_{\text{BL Lacs}}$ ；第 8-11 列为 5BZCAT、3FGL、4FGL 和 Kang 等[15-16]对 BCUs 的分类；第 12 列为 $p_0=0.8$ 时，4 个分类器共同决定的 BZUs 类别。

最终，本文得到 33 个 BL Lacs 候选体和 119 个 FSRQs 候选体，两者比例与 Kang 等[15-16]不一致，主要是因为:1)BZUs 是从不同文献收集来的，例如，f_R 来自 NVSS（The Nrao VLA Sky Survey）[22]，$m_R^{'}$ 来自 USNO-B（The USNO-B Catalog）[23]和 SDSS DR10(The Sloan Digital Sky Survey quasar catalog: tenth data release)[24]。而 BCUs 是来源于 Fermi 观测，所以 BZUs 和 BCUs 中 BL Lacs 和 FSRQs 的分布不同；2)本文与 Kang 等[15-16]使用的判别概率阈值不同。Kang 等[15-16]取 0.5 为判别概率阈值，而本文取 0.8，这也使得本文对 BL Lacs 候选体的判定较为严苛，导致本文所得 BL Lacs 候选体较少；3）本文与 Kang 等[15-16]使用的用于分类的参数不同，本文使用的参数为 5BZCAT 表提供，而 Kang 等[15-16]使用的参数则来自费米表。相较于 Kang 等[15-16],本文对每个分类器中重要的 HPs(包括 p_0)进行了较细致的筛选(GS)。本文中的分类准确率最高达到 94.81%，略优于 Kang 等[15-16]的 91.6%和 92.9%。此外,Kang 等[15]的工作表明,训练集和测试集的划分比例不同,得到的分类结果也会不同,而本文只考虑 7：3 这个比例，因此本文的结果可能有一定的偏向性。本文注意到 227 个 BZUs 中,有部分在 3FGL 和 4FGL 中被分类为非耀变体源,而本文分类时只考虑 BL Lacs 和 FSRQs 这两个耀变体的子类别，对 BZUs 的分类是否该考虑更多的候选类别，而不仅限于 BL Lacs 和 FSRQs，这个问题超出了本文的研究范围。应当指出，判定一个耀变体是 BL Lacs 还是 FSRQs，最准确的方法仍然是光学光谱测量，ML 方法可作为高效的替代方法，为后续可能的光学光谱测量提供可信度较高的候选体。

表 3 4种分类器与其他文献的不匹配源

Table 3 Mismatched sources of 4 classifiers

mismatched	total	model	$p_0=0.5$	$p_0=0.6$	$p_0=0.7$	$p_0=0.8$	$p_0=0.9$	$p_0=0.95$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
3FGL	38	SVM	14	12	9	9	9	9
		RF	17	16	13	13	16	17
		MLP	11	12	13	14	15	17
		EM	5	5	7	8	8	7
4FGL	48	SVM	7	6	6	6	6	5

		RF	12	11	11	11	11	10
		MLP	10	11	11	11	11	12
		EM	10	11	11	10	10	9
Kang[15]	26	SVM	10	10	11	11	10	10
		RF	13	12	14	14	13	12
		MLP	10	11	12	12	11	12
		EM	8	8	9	9	9	9
Kang[16]	20	SVM	8	10	9	8	7	7
		RF	12	13	12	13	12	12
		MLP	10	10	11	11	10	11
		EM	12	11	10	10	9	8

表 4　联合4种分类器后与其他文献的不匹配源

Table 4　Mismatched sources of combining 4 classifiers

mismatched	total	$p_0=0.5$	$p_0=0.6$	$p_0=0.7$	$p_0=0.8$	$p_0=0.9$	$p_0=0.95$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
3FGL	38	5	5	6	5	4	4
4FGL	48	7	6	8	7	6	6
Kang[15]	26	5	5	8	8	8	8
Kang[16]	20	2	2	4	5	5	4
unks		88	91	76	75	82	75

表 5　227个BZUs的分类结果与其他文献的分类

Table 5　Classification results of 227 bzus and other literatures

5BZCAT name	3FGL name	4FGL name	SVM	RF	MLP	EM	3LAC Class	4FGL Class	Kang[15]	Kang[16]	Class
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
5BZU J1312+4828	3FGL J1312.7+4828	4FGL J1312.6+4828	0.2476	0.417	0.0038	0.1812	agn	bcu			fsrq
5BZU J0040+4050	3FGL J0040.3+4049	4FGL J0040.3+4050	0.5685	0.997	0.9823	0.9809	bcu	bll	fsrq		unks

5BZU J0049-4457	3FGL J0050.0-4458	4FGL J0049.6-4500	0.9096	0.909	0.9941	0.9668	bcu	bcu	bll	fsrq	bll
...	...	...	...	...	...	...	...	...	...	...	...
5BZU J2352+3947			0.6125	0.001	0.094	0.0481					fsrq
5BZU J2354-0405			0.4935	0.065	0.0335	0.0437					fsrq
5BZU J2354-4106			0.561	0.004	0.032	0.0866					fsrq

表注：完整表格请参考附件

4 总结

本文以 5BZCAT 为主要样本，结合 NED 数据，选取红移，多波段有效谱指数，多波段流量/流量密度等 8 个参数，用 SVM，RF，MLP，EM 这 4 种 ML 分类器对 5BZCAT 中的 227 个 BZUs 进行分类，用特征工程和网格搜索分别筛选最优的特征和 HPs，提升分类准确率。并与其他文献的分类结果进行了比较，通过将判别概率阈值 p_0 设为 0.8，并同时考虑 4 种分类器的预测类别，进一步减少了与其他文献不匹配的源。本文的分类结果表明，BL Lacs 和 FSRQs 在 8 参数空间中是可区分的，最终得到 33 个 BL Lacs 候选体和 119 个 FSRQs 候选体。

致谢：本工作得到了国家自然科学基金(11733001、U1531245)、广东省自然科学基金(2017A030313011；2019B030302001)、广东省和广州市重点学科的资助，特此感谢！

参考文献：

[1] Padovani P，Giommi P. The Connection between X-Ray and Radio-selected BL Lacertae Objects [J]. The Astrophysical Journal，1995，444: 567-581.

[2] Aller M F，Aller H D，Hughes P A. Pearson-Readhead Survey Sources: Properties of the Centimeter-Wavelength Flux and Polarization of a Complete Radio Sample [J]. The Astrophysical Journal，1992，399: 16-28.

[3] Fan J H，Huang Y，He T M，et al. Radio Variability and Relativistic Beaming Effect for Blazars [J]. Publications of the Astronomical Society of Japan，2009，61: 639-643.

[4] Lin C，Fan J H，Xiao H B. The intrinsic γ-ray emissions of Fermi blazars [J]. Research in Astronomy and Astrophysics，2017，17: 66-80.

[5] Abdo A A，Ackermann M，Agudo I，et al. The Spectral Energy Distribution of Fermi Bright Blazars [J]. The Astrophysical Journal，2010，716: 30-70.

[6] Fan J H，Yang J H，Liu Y，et al. The Spectral Energy Distributions of Fermi Blazars [J]. The Astrophysical Journal Supplement Series，2016，226: 20-38.

[7] Massaro E，Maselli A，Leto C，et al. The 5th edition of the Roma-BZCAT. A short presentation [J]. Astrophysics and Space Science，2015，357: 75.

[8] Acero F，Ackermann M，Ajello M，et al. Fermi Large Area Telescope Third Source Catalog [J]. The Astrophysical Journal Supplement Series，2015，218: 23-64.

[9] Abdollahi S，Acero F，Ackermann M，et al. Fermi Large Area Telescope Fourth Source Catalog [J]. The Astrophysical Journal Supplement Series，2020，ApJS，247: 33.

[10] Ball N M，Brunner R J. Data Mining and Machine Learning in Astronomy [J]. International Journal of Modern Physics D. 2010，19: 1049-1106.

[11] Hobson M，Graff P，Feroz F，et al. Machine-learning in astronomy[J]. Proceedings of the International Astronomical Union，2014，10: 279-287.

[12] Baron D. Machine Learning in Astronomy: a practical overview. 2019，arXiv: 1904.07248.

[13] Hassan T，Mirabal N，Contreras J L，et al. Gamma-ray active galactic nucleus type through machine-learning algorithms [J]. Monthly Notices of the Royal Astronomical Society，2013，428:220-225.

[14] Lefaucheur J，Pita S. Research and characterisation of blazar candidates among the Fermi/LAT 3FGL catalogue using multivariate classifications [J]. Astronomy and Astrophysics，2017，602: 86-98.

[15] Kang S J，Fan J H，Mao W M，et al. Evaluating the Optical Classification of Fermi BCUs Using Machine Learning [J]. The Astrophysical Journal，2019，872: 189-200.

[16] Kang，S. J，Li E，Ou W T，et al. Evaluating the Classification of Fermi BCUs from the 4FGL Catalog Using Machine Learning [J]. The Astrophysical Journal，2020，887: 134-144.

[17] Chiaro G，Salvetti D，La Mura G，et al. Blazar flaring patterns (B-FlaP) classifying blazar candidate of uncertain type in the third Fermi-LAT catalogue by artificial neural networks [J]. Monthly Notices of the Royal Astronomical Society，2016，462: 3180-3195.

[18] Kovačević M，Chiaro G，Cutini S，et al. Optimizing neural network techniques in classifying Fermi-LAT gamma-ray sources [J]. Monthly Notices of the Royal Astronomical Society，2019，490: 4770-4777.

[19] Kovačević M，Chiaro G，Cutini S，et al. Classification of blazar candidates of uncertain type from the Fermi LAT 8-yr source catalogue with an artificial neural network [J]. Monthly Notices of the Royal Astronomical Society，2020，493:1926-1935.

[20] Ackermann M，Ajello M，Atwood W B，et al. The Third Catalog of Active Galactic Nuclei Detected by the Fermi Large Area Telescope [J]. The Astrophysical Journal，2015，810: 14-48.

[21] Ral Garreta. Learning scikit-learn: Machine Learning in Python[M]. Packt Publishing, 2013.

[22] Condon J J, Cotton W D, Greisen E W, et al. The NRAO VLA Sky Survey[C].Bulletin of the American Astronomical Society, 1993, 25:1389.

[23] Monet D G, Levine S E, Canzian B, et al. The USNO-B Catalog[J]. The Astronomical Journal, 2003, 125: 984-993.

[24] P âris I, Petitjean P, Aubourg É, et al. The Sloan Digital Sky Survey quasar catalog: tenth data release[J]. Astronomy and Astrophysics, 2013, 563: 54-15.

Classication for BZUs in 5BZCAT

J. T. Zhu，J. H. Fan，J. T. Cai

(Center for Astrophysics，Guangzhou University，Guangzhou 510006，China)

(Astronomy Science and Technology Research Laboratory of Department of Education of Guangdong Province，Guangzhou 510006，China)

(Key Laboratory for Astronomical Observation and Technology of Guangzhou，Guangzhou 510006，China)

Abstract: In order to evaluate the potential optical classification of 227 BZUs in 5BZCAT, we divided the BZUs into BL lac candidates and FSRQ candidates by four machine learning methods: support vector machine (SVM), random forest (RF), ensemble learning (EM) and multi-layer perceptron (MLP). And the classification accuracy is improved by feature engineering and grid search. By combining the classification results of four classifiers and setting the threshold of discrimination probability to 0.8, we get 33 BL lacs candidates and 119 FSRQs candidates.

Key words:BL Lac objects; Flat Spectrum Radio Quasars; machine learning; classification